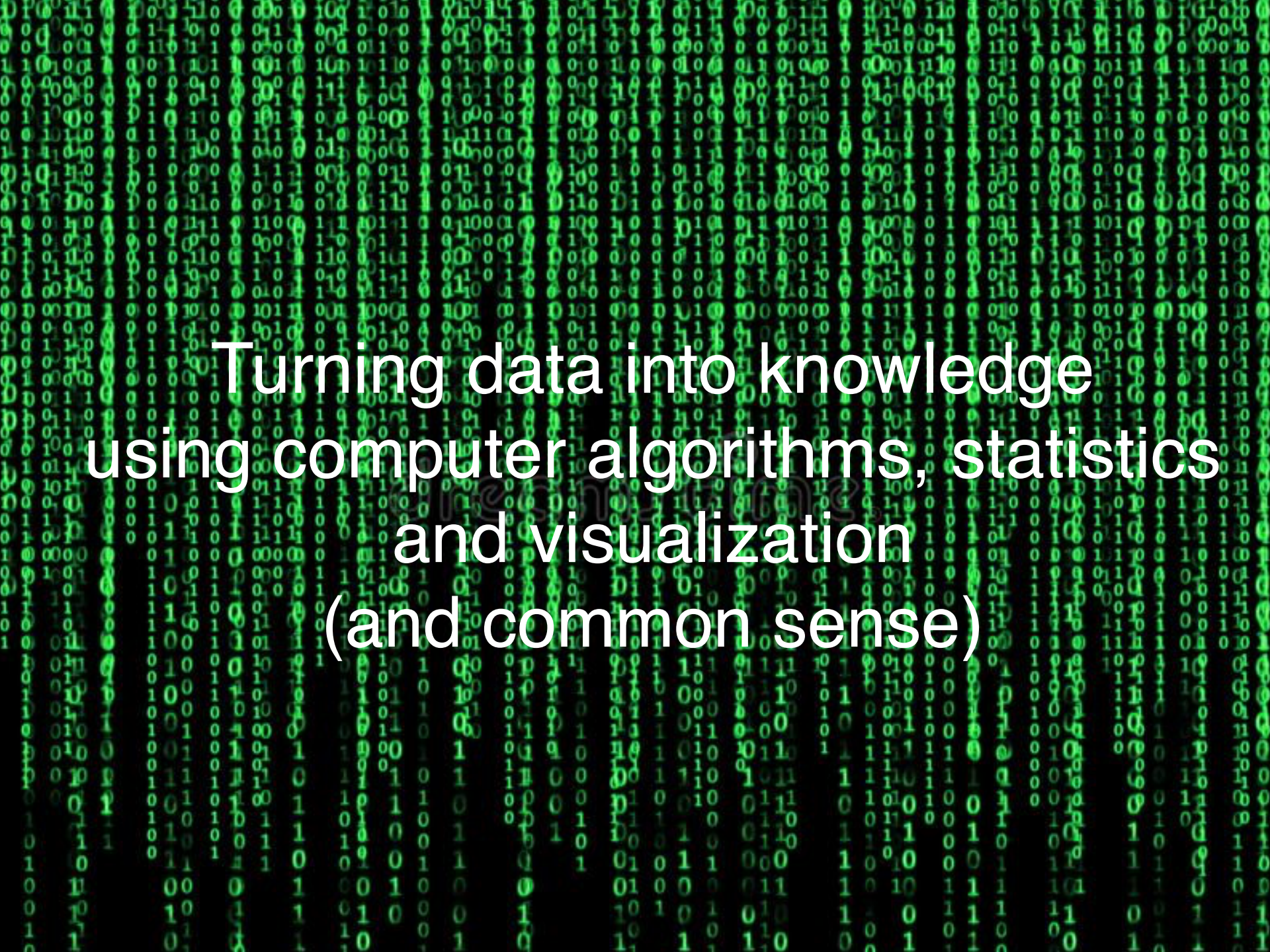
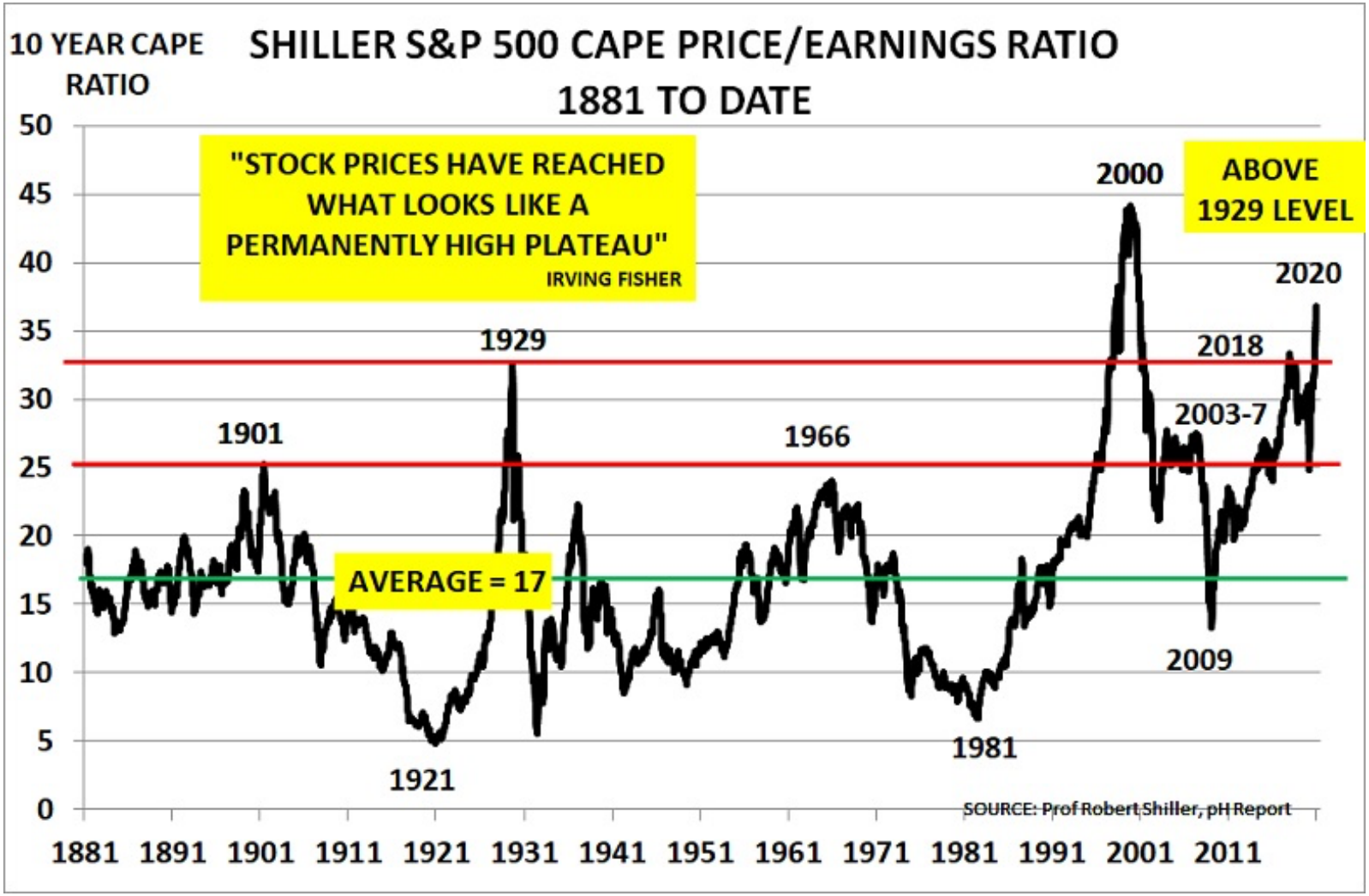
Data Analysis and Machine Learning using Python

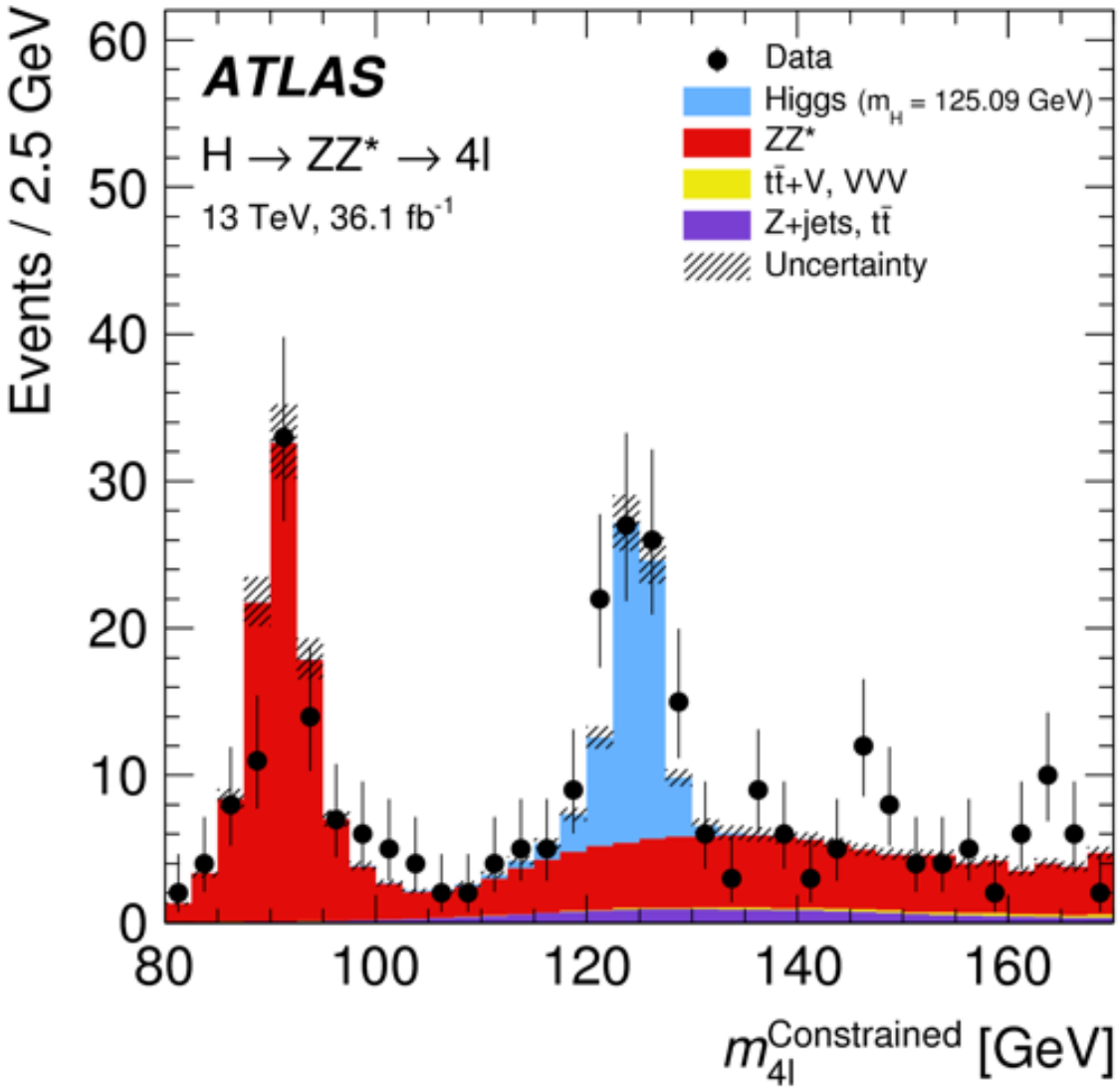
Lecture 4: Visualizing data

March 30 2024


Turning data into knowledge
using computer algorithms, statistics
and visualization
(and common sense)



Describe patterns and correlations in financial markets
(and, maybe, predict future)

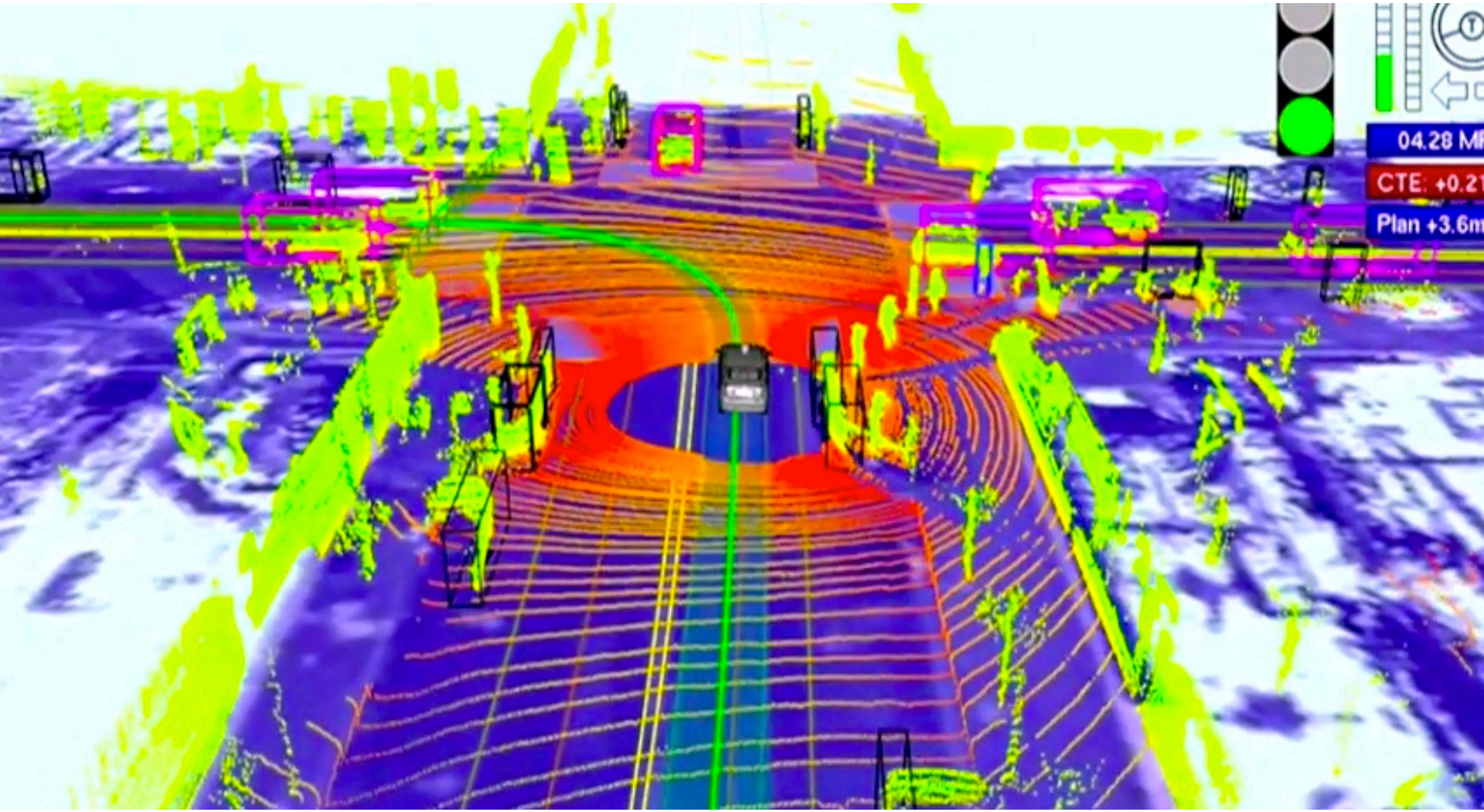


Discover

(and, definitely, predict where things will move to)

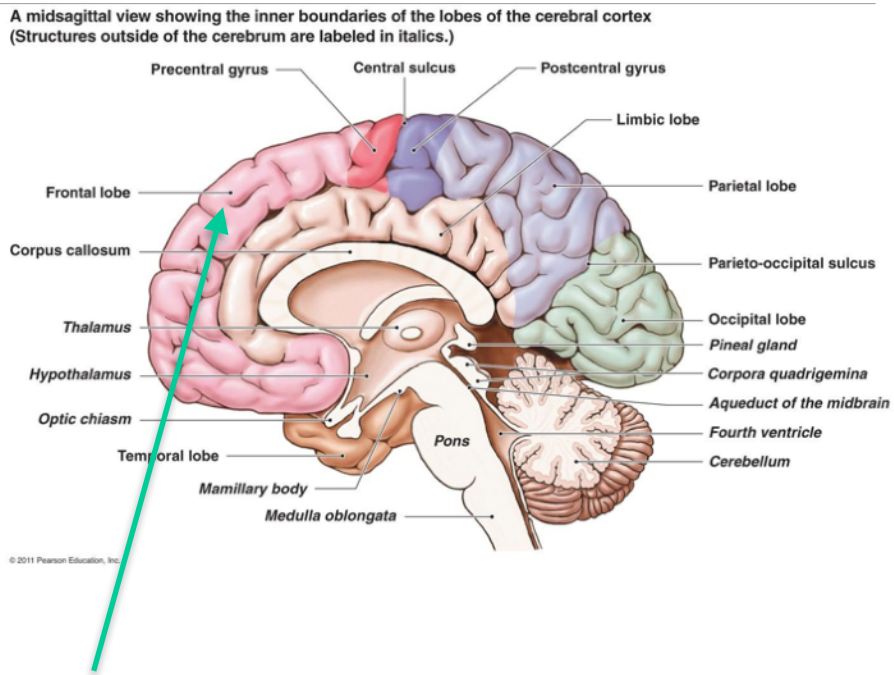
Different example

- First two examples show a plot communicating the final result of analysis in human-understandable form
- A different aspect of visualization is quality control - visualize the state or result of a calculation to enable human cross-check
  - e.g., what does a self-driving car “see”?



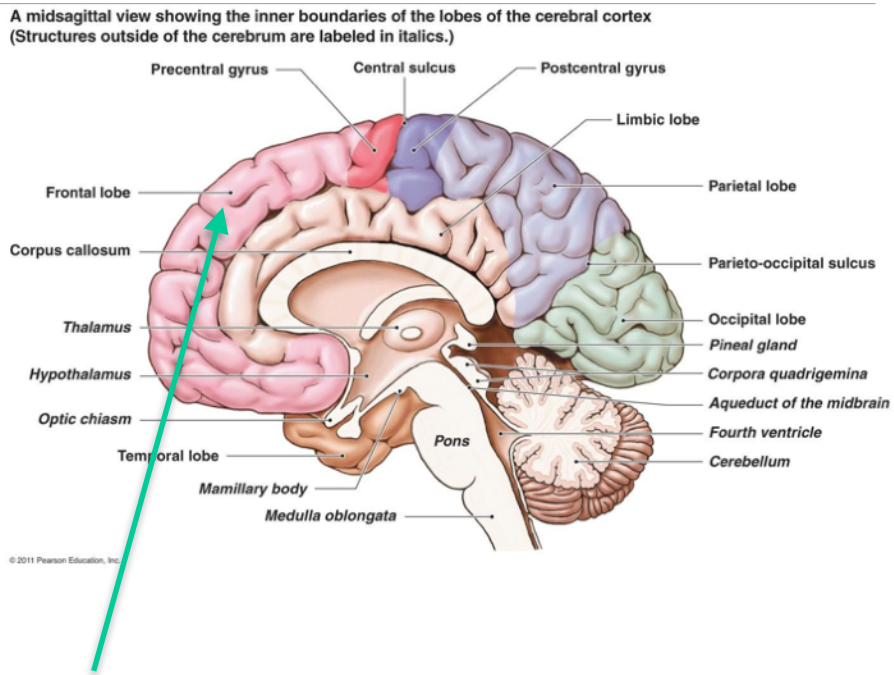
Make a dynamic model of surroundings
(and, definitely, predict where things will move to)

Why visualization?



How to get information from here to there?

Why visualization?



How to get information from there into this?

"**More than 50 percent** of the cortex, the surface of the brain, is devoted to processing visual information," points out Williams, the William G. Allyn Professor of Medical Optics. "Understanding how vision works may be a key to understanding how the brain as a whole works."

Human vision

- High bandwidth
- Fast, parallel processing
- Subconscious (pre-attentive) processing
- Powerful pattern recognition
  - e.g., facial recognition of people with masks, looking in the wrong direction, wearing glasses…
- Direct connection to our model of reality
  - people “think visually”
  - many great scientists were visual thinkers
- → use vision/visualization to transfer information

Visualization

- Facilitate cognition through computer supported/ generated visual representations
  - Information visualization
  - Scientific visualization
  - Virtual/augmented reality
-

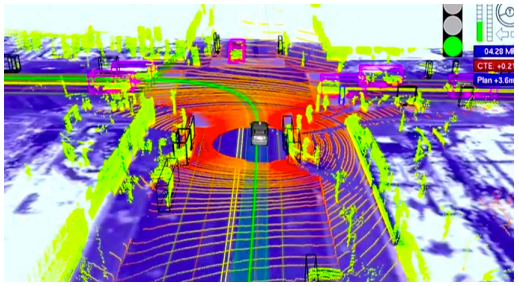
Visualization

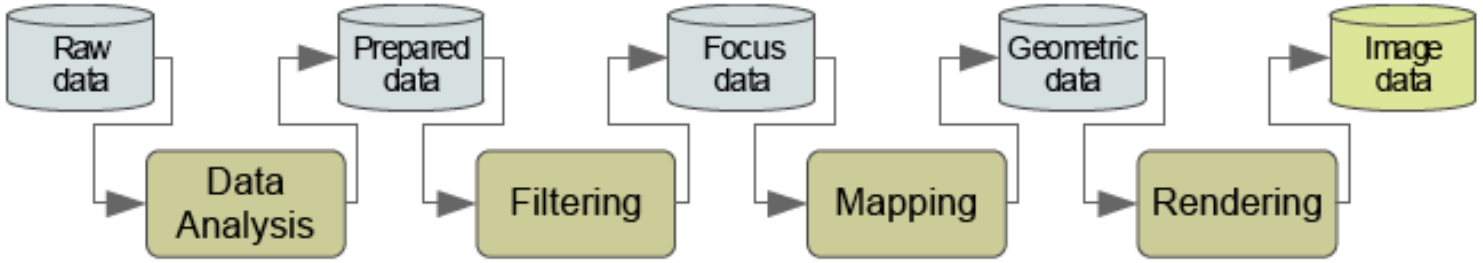
- Facilitate cognition through computer supported/ generated visual representations
  - Information visualization
  - Scientific visualization
  - ~~Virtual/augmented reality~~
- Principles that allow for efficient cognition
- Tools for visualization in Python - Matplotlib

Basic principles

- How do we present graphics to allow recognition of
  - Relationships, correlations, trends
  - illustrate concepts, results and conclusions
- in a way that is statistically and scientifically sound and unbiased?
- Beware of statistical (and graphical) sins!

Visualization pipeline





- What are principles for sound visual representation?
- Note: can’t separate visualization from statistical and analytical soundness of data and processing

The power of visualization

- Relationship between college degree % and income?
- Any outliers/anomalies?

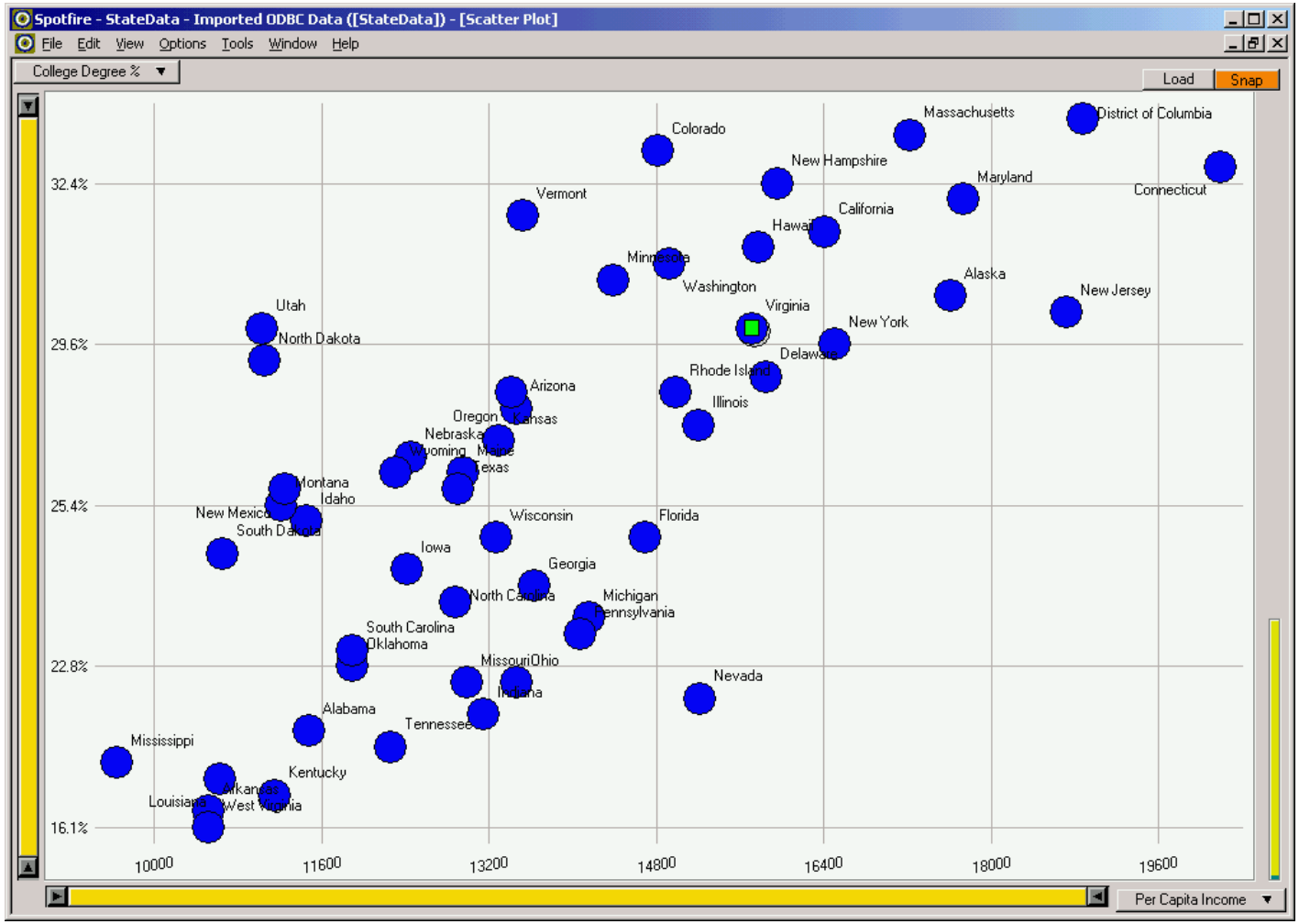
Table - StateData ()

Load | Snap

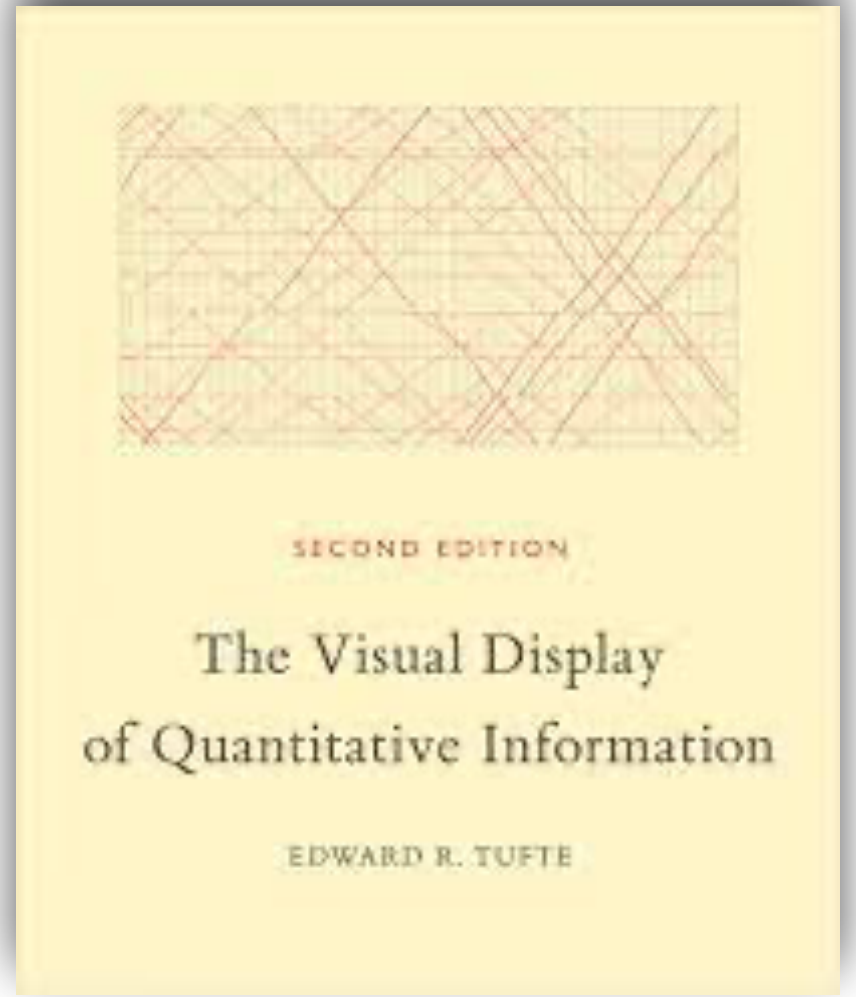
State	College Degree %	Per Capita Income
Alabama	20.6%	11486
Alaska	30.3%	17610
Arizona	27.1%	13461
Arkansas	17.0%	10520
California	31.3%	16409
Colorado	33.9%	14821
Connecticut	33.8%	20189
Delaware	27.9%	15854
District of Columbia	36.4%	18881
Florida	24.9%	14698
Georgia	24.3%	13631
Hawaii	31.2%	15770
Idaho	25.2%	11457
Illinois	26.8%	15201
Indiana	20.9%	13149
Iowa	24.5%	12422
Kansas	26.5%	13300
Kentucky	17.7%	11153
Louisiana	19.4%	10635
Maine	25.7%	12957
Maryland	31.7%	17730
Massachusetts	34.5%	17224
Michigan	24.1%	14154
Minnesota	30.4%	14389
Mississippi	19.9%	9648
Missouri	22.3%	12989
Montana	25.4%	11213
Nebraska	26.0%	12452
Nevada	21.5%	15214
New Hampshire	32.4%	15959
New Jersey	30.1%	18714
New Mexico	25.5%	11246
New York	29.6%	16501
North Carolina	24.2%	12885
North Dakota	28.1%	11051
Ohio	22.3%	13461
Oklahoma	22.8%	11893
Oregon	27.5%	13418
Pennsylvania	23.2%	14068
Rhode Island	27.5%	14981
South Carolina	23.0%	11897
South Dakota	24.6%	10661
Tennessee	20.1%	12255
Texas	25.5%	12904
Utah	30.0%	11029
Vermont	31.5%	13527
Virginia	30.0%	15713
Washington	30.9%	14923
West Virginia	16.1%	10520
Wisconsin	24.9%	13276
Wyoming	25.7%	12311

The power of visualization

- Relationship between college degree % and income?
- Any outliers/anomalies?



Representing information

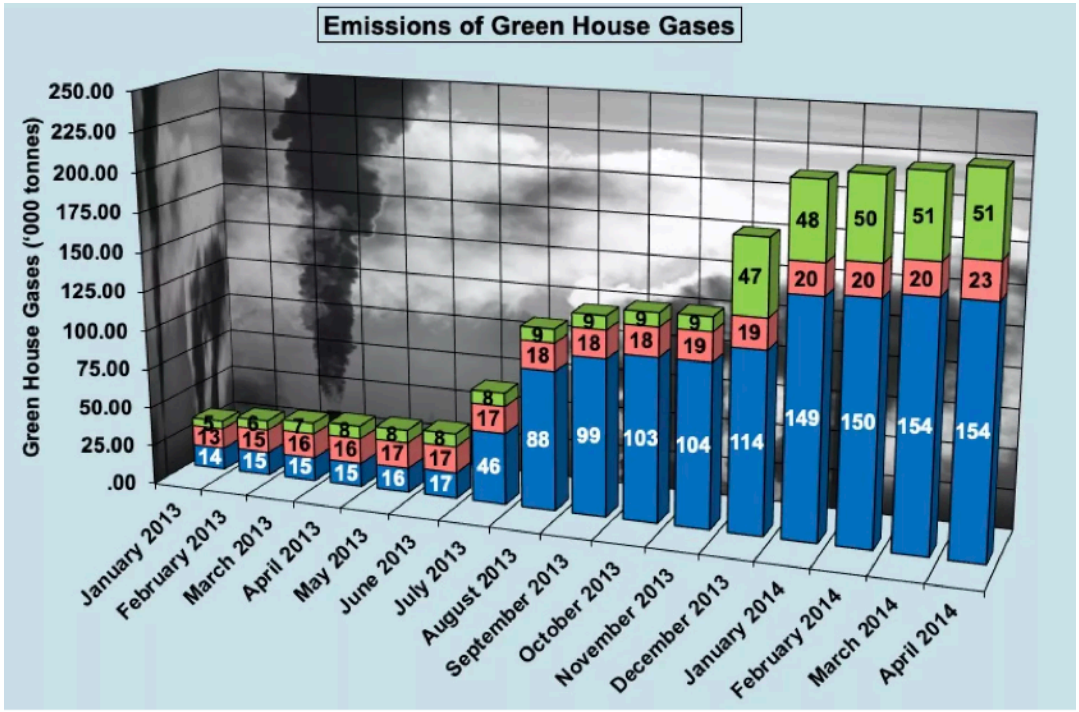


Great text explaining many of the principles behind good (and bad) visual representation of information

Edward Tufte (Yale)

"Above all else show the data"

- Tufte describes the "data-ink" ratio, the proportion of the "ink" used to present data vs the total amount of ink in the graph
- E.g., don't do this:

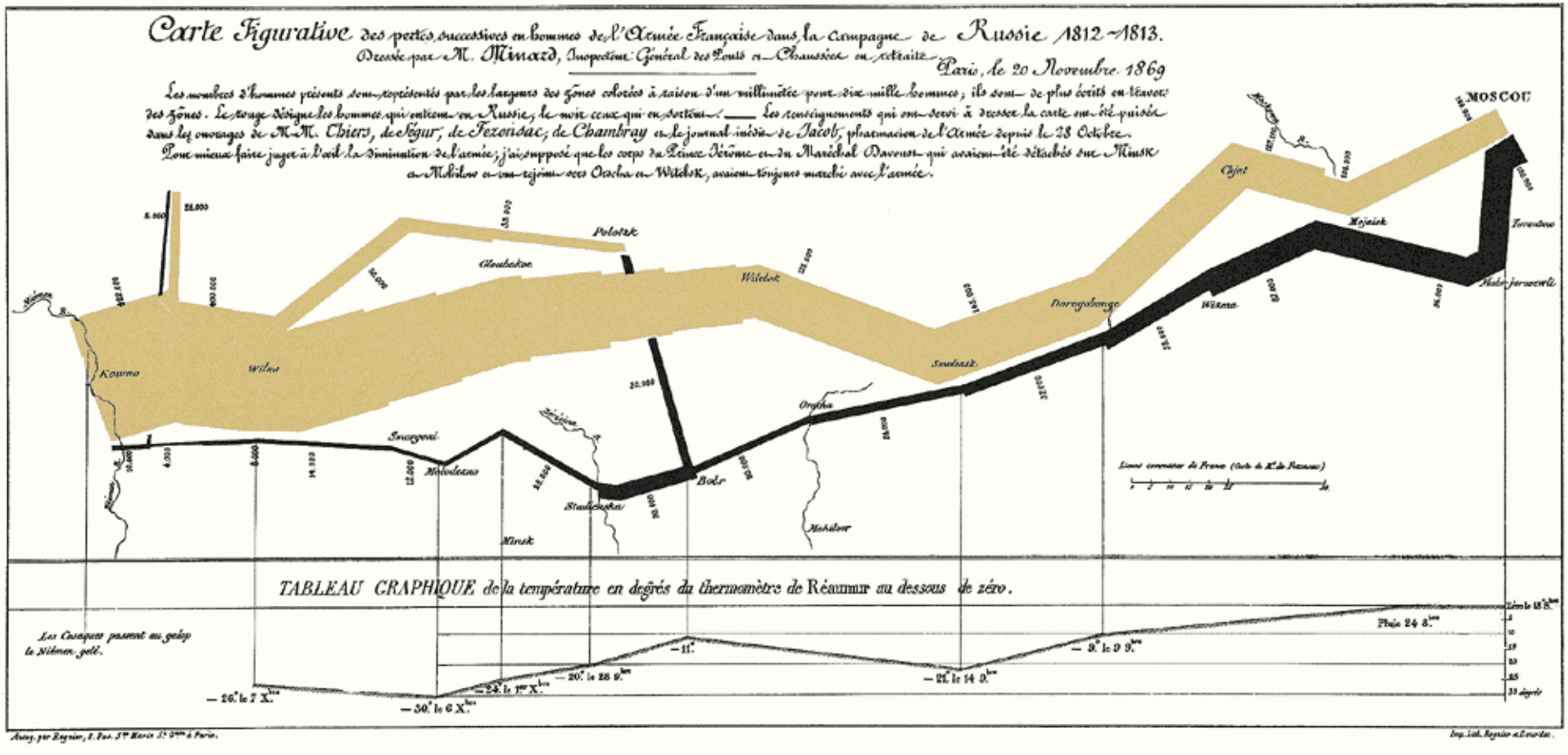


Background image and pseudo-3D perspective add no information, but make it harder to process data

Also true when designing slides - background graphics are most often distracting without adding information

A classic (good) example

Napoleon's march on Moscow



(Nearly) every drop of ink on this graph conveys important information
- position, color and thickness of lines all have meaning

How to do this?

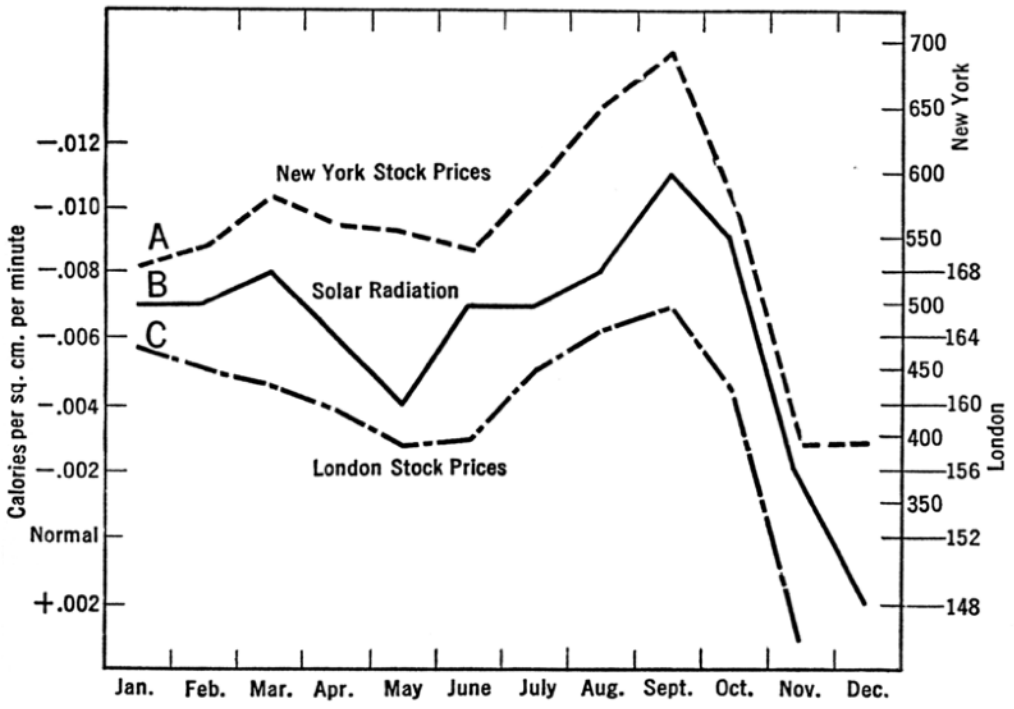
"Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficacy"

- Show the data!
- Don't distort the data
- Encourage eye to explore the data
- Make large data sets compact and coherent
- Present multiple levels of detail
- Connect to statistical and verbal description of information
- Understand the purpose of the graph

How not to do this

- Arbitrary colors convey no information
- Tiny, sideways names hard to read for no good reason
- Big graph for little information (data-ink ratio)

Of course, statistical graphics, just like statistical calculations, are only as good as what goes into them. An ill-specified or preposterous model or a puny data set cannot be rescued by a graphic (or by calculation), no matter how clever or fancy. A silly theory means a silly graphic:



SOLAR RADIATION AND STOCK PRICES

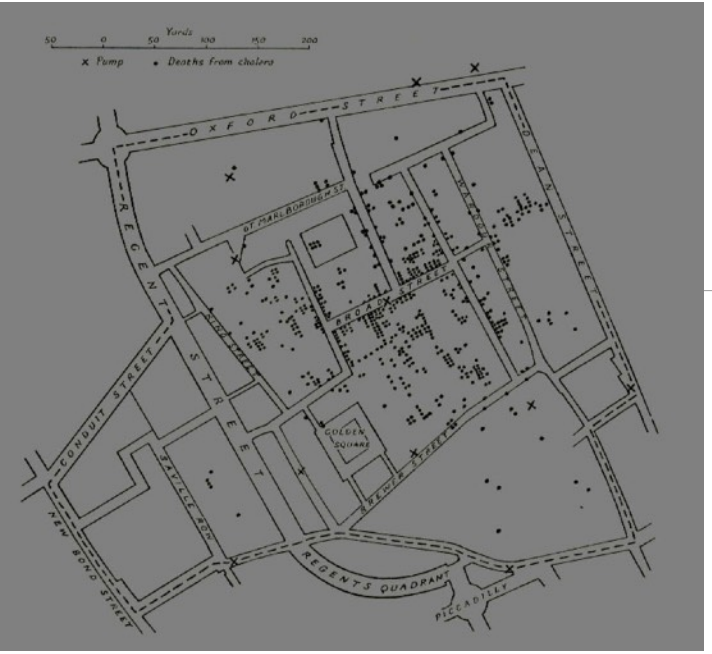
A. New York stock prices (Barron's average). B. Solar Radiation, inverted, and C. London stock prices, all by months, 1929 (after Garcia-Mata and Shaffner).

Just because one can plot unrelated things next to each other, doesn't mean one should

If you compare enough random things to each other, you will find some that are statistically correlated

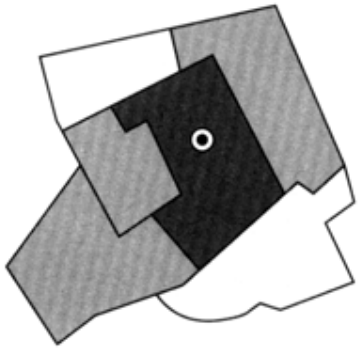
That does not imply causation

Related to "look elsewhere" effect

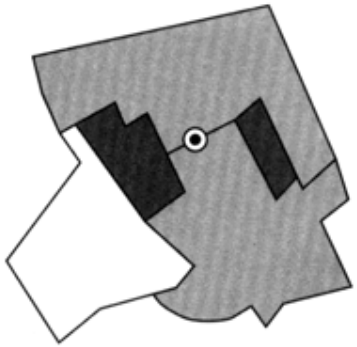


If the statistical power of your dataset is poor, arbitrary changes in plotting can lead to opposite conclusions

Beware of sub-conscious biases!

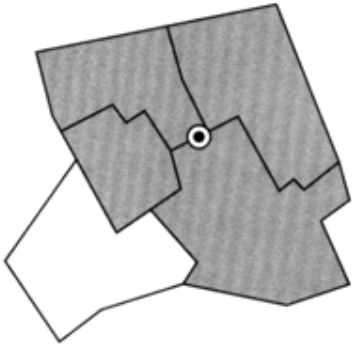


In this aggregation of individual deaths into six areas, the greatest number is concentrated at the Broad Street pump.



In this aggregation of the deaths, the two areas with the most deaths do not even include the infected pump!

[18] Mark Monmonier, *How to Lie with Maps* (Chicago, 1991), pp. 142–143.

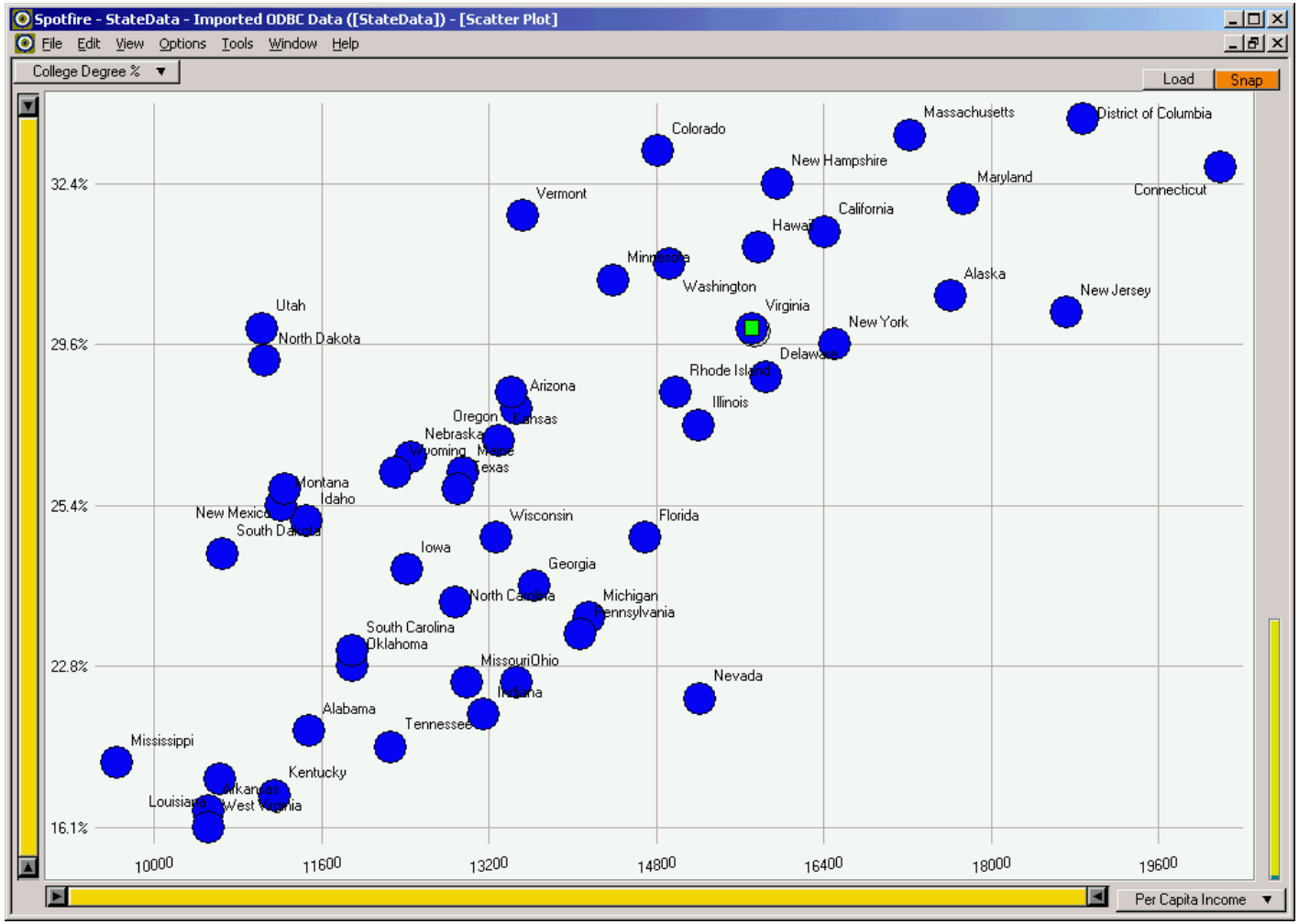


Using different geographic subdivisions, the cholera numbers are nearly the same in four of the five areas.

Perception and interpretation

- Human brain is extremely good at pattern recognition
  - essential survival skill
  - much of it subconscious (pre-attentive)
- Leverage these capabilities to reduce cognitive overhead in absorbing information
- But beware, humans sometimes see patterns where there are none!

Subconscious image processing - correlation is evident without even trying



Perception and interpretation

- Visual processing (subconscious) can instantly identify trends, structures, patterns in displayed data
- Followed by interpretation (conscious) of meaning of patterns
- When designing graphics you need to anticipate and facilitate the subconscious and conscious processing of the displayed information
- Don’t follow a principle of “it was hard to analyze, so it should be hard to understand”
- What do you want the audience to see?
- What do you want them to understand?

Some hints

- Use size, color, *ink* to emphasize what is important
- Make text easy to read
- Don’t hide your data - should fill the graph!
- Be consistent
  - consistent meaning of color
  - consistent meaning of symbols (markers)
  - consistent order of elements
- Use as little text as necessary, but not less
- Don’t use plots as excuse/filler, but expect that audience will try to understand what is shown!

Color

12345678901234567890123466789O

Color

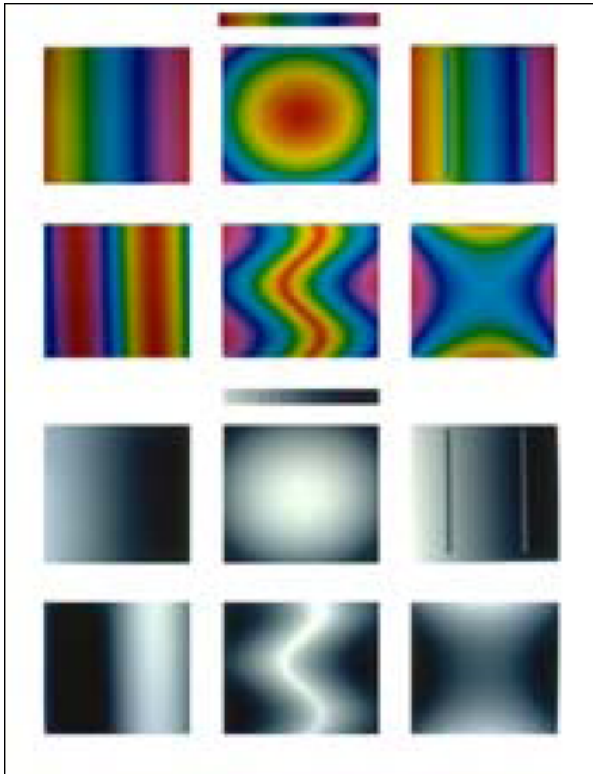
12345678901234567890123467890

Color provides powerful pre-attentive visual aide when used appropriately

Color

12345678901234567890123467890

Color provides powerful pre-attentive visual aide when used appropriately



- Color scales however are not pre-attentive
- require conscious effort for detailed interpretation

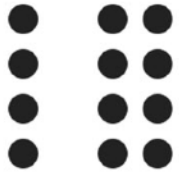
Gestalt laws

Gestalt Principles

UX CHEAT.com

Good Figure

Objects groupped together tend to be perceived as a single figure. Tendency to simplify.



Proximity

Objects tend to be grouped together if they are close to each other.

Similarity

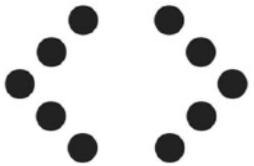
Objects tend to be grouped together if they are similar.

Continuation

When there is an intersection between two or more objects, people tend to perceive each object as a single uninterrupted object.

Closure

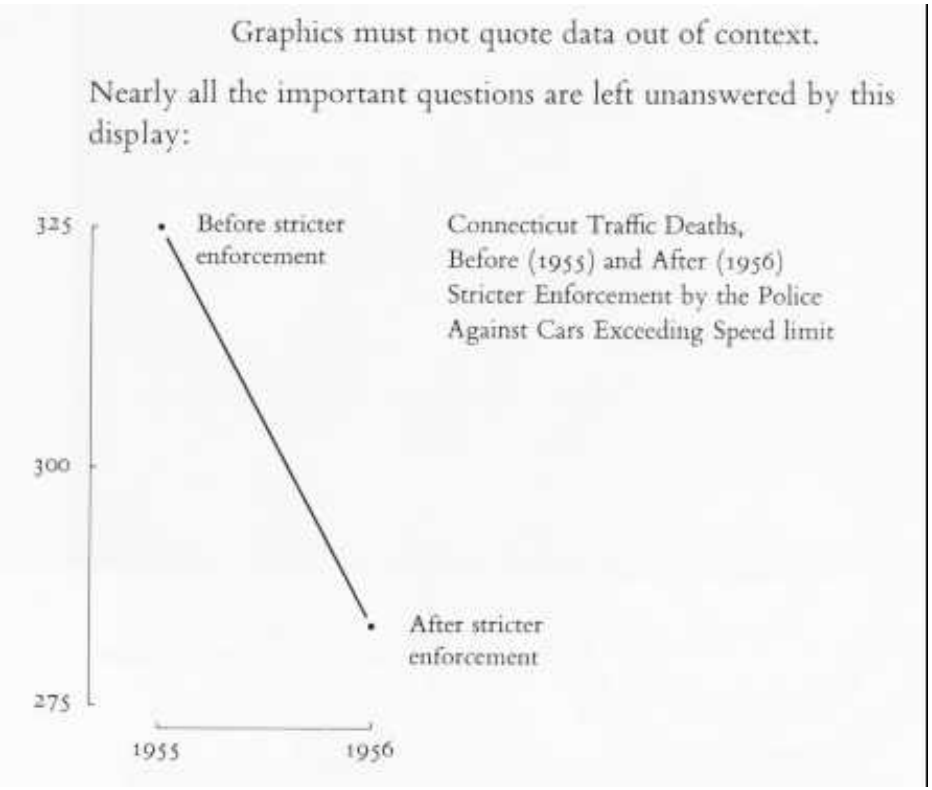
Visual connection or continuity between sets of elements which do not actually touch each other in a composition.



Symmetry

The object tend to be perceived as symmetrical shapes that form around their center.

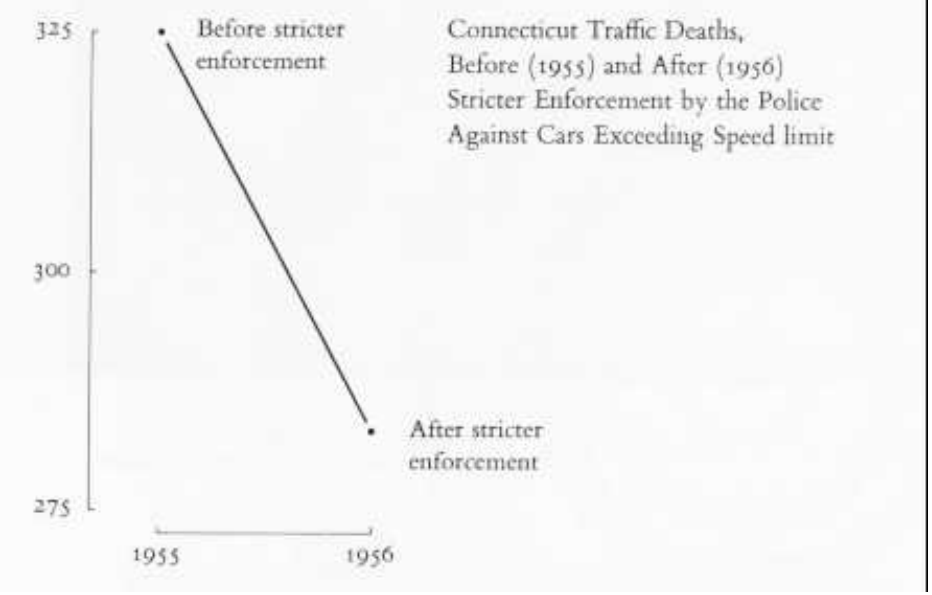
Lying with plots



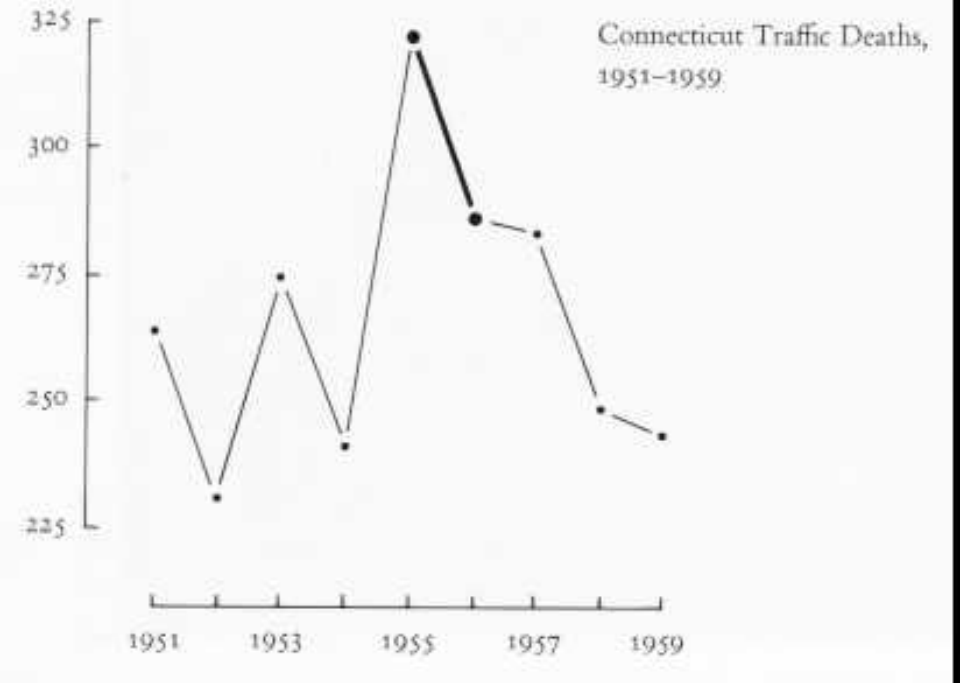
Lying with plots

Graphics must not quote data out of context.

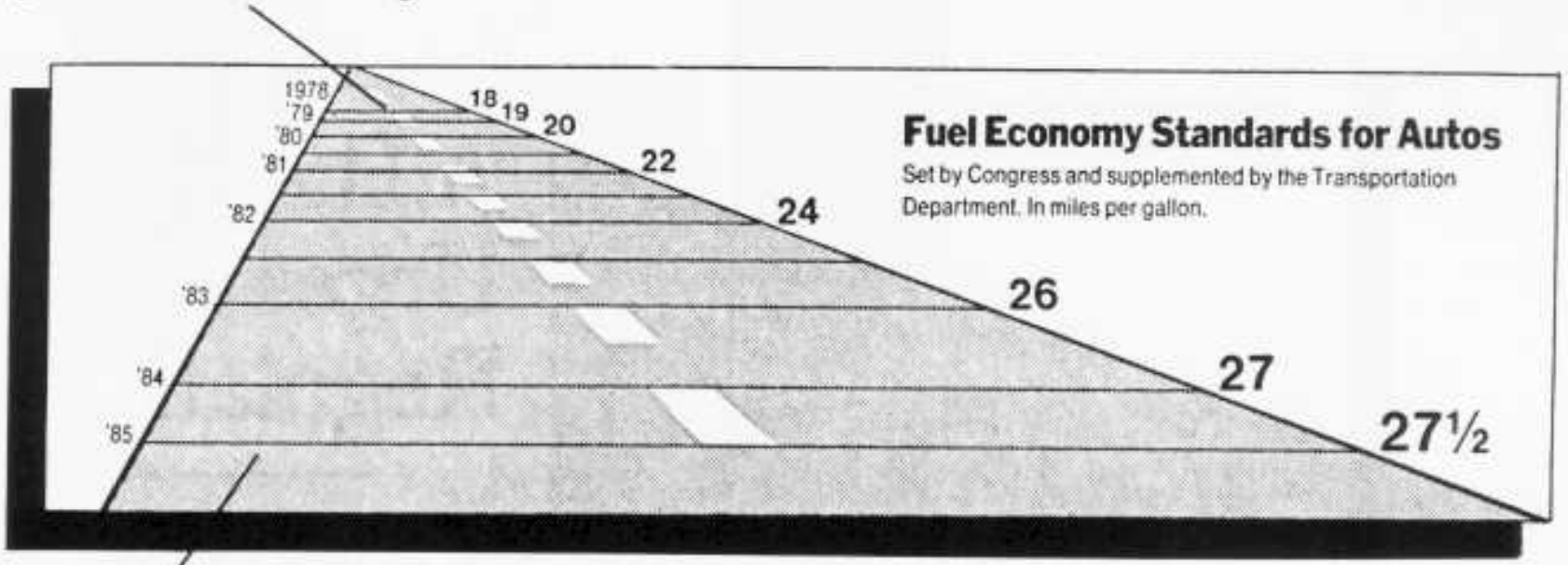
Nearly all the important questions are left unanswered by this display:



A few more data points add immensely to the account:



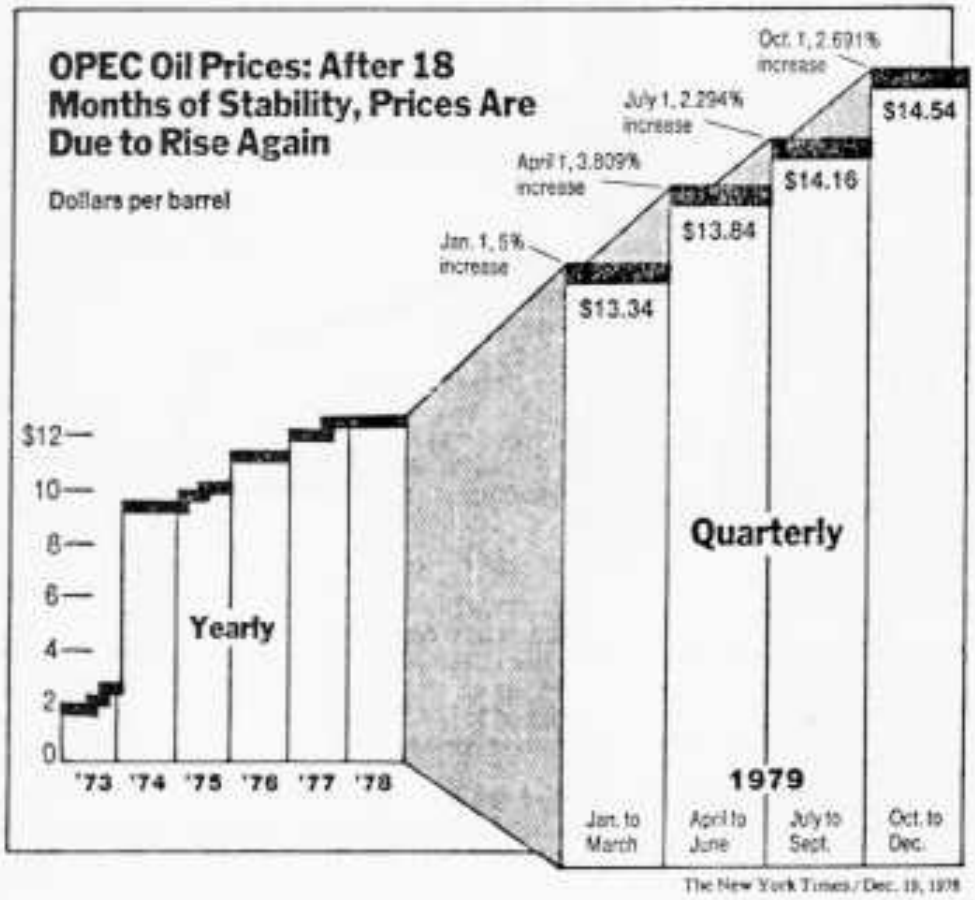
Lying with plots



New York Times, August 9, 1978, p. D-2.

Lying with plots

Design variation corrupts this display:



New York Times, December 19, 1978, p. D-7.

Homework, part 1

- Using the data file data_HW1.txt, plot the x,y values of neighboring columns against each other using, e.g., plt.plot()
- This will give 4 plots. Plot them as 2x2 subplots on one figure

Homework, part 2

- Using the TopRight_20230803.txt file, make separate 1D histograms of the data in the following columns:
  - timestamp
  - SiPM
  - Temperature
  - Pressure
- This will give 4 plots. Plot them as 2x2 subplots on one figure
- Make a 2D histogram (using hist2D) of timestamp vs Pressure
- Make sure all plots have legible axis labeling and axis titles and generally look “nice”